Advanced methods in big data

Assessment

S. Hué, P. Michel

To apply the knowledge you have acquired in the three courses of the "Advanced methods in big data" Teaching Unit (=UE), you will produce a report in groups of 2 or 3 people. For practical reasons, we strongly encourage you to work with students enrolled in the same program as you (students enrolled in the work-link training should not work with students not enrolled in this program). You will defend your work in an oral presentation. Your work will be based on a dataset of your choice. Although the choice of the subject on which you will work is at your discretion, it shall include some elements common to all three courses of the teaching unit, in addition to dealing with an economic issue. To produce your report (no more than 35 pages), you will draw on the knowledge and skills acquired in each of the three courses involved in the assessment.

From a formal point of view, your report should include the following elements:

• A detailed introduction specifying your objective and your approach to address it. The study must be motivated by your knowledge as an economist. The introduction should allow the reader to answer the following points:

- what you are focusing on
- why it is relevant (think as economists here)
- what is already known in the literature (and why it needs attention)
- how you address the question (here, using an empirical analysis)
- what you specifically did (strategy and methods, in a nutshell)
- what you found (highlight of your key results)

Ideally, you should be able to cite in your introduction academic articles that propose a topic related to the one you are studying.

• A detailed section: "Materials and Methods" presenting first the data used, the ways to collect and format them; and a detailed description of the methods used in your analysis: theory elements on the models used, empirical strategy (model selection, cross validation, grid search, etc.). Do not forget to cite your sources.
Note: Your work should not be a catalogue of estimation methods. It is best to limit yourself to a small number of methods, which you can present properly.

• A section presenting the **results** of your study. This will be divided into two sub-sections:

- first, the results of an **exploratory analysis**, in which you will provide a summary of the data (visual and statistical). Please note that this part is not to be neglected. You will take care to make an effort in the choice of tools to describe your data. Be sure to include relevant comments from reading each graph and table.
- second, the results of the **predictive analysis** (classification problem, or regression problem).

• A **conclusion** recalling your major results and proposing possible future work and perspectives.

Evaluation criteria

To enable you to propose a substantial amount of work rather than three modest projects, we ask you to work on a single project. This project gives rise to three grades, one for each component of the following Teaching Unit (=ECUE):

- Méthodes de réduction de l'information
- Méthodes de prévision
- Machine learning et statistical learning

Each teacher will evaluate you on points related to the course they have given you.

Warning: it may be tempting to divide the work within the group in order to create task specialisation. It would be smarter to share your knowledge about the whole project so that each of you can progress. Do not forget that projects are excellent ways of putting learning by doing into practice.

Written assignment

Your report will be evaluated considering the following points:

- Data exploration and preprocessing.
- Implementation of dimension reduction techniques and understanding of the underlying theory.
- Application and understanding of the methods used (supervised and / or unsupervised).
- Computational aspect (vectorization of computer code – if needed–, explanation of the functioning of the R or Python libraries used).
- The quality of the report formatting and compliance with instructions (35 pages maximum).

Finally, do not forget to use your economic knowledge to bring a more in-depth view than that of a data manipulator. You are neither data engineers nor data scientists, you are economic data scientists. Also consider specifying the interest that the use of machine learning techniques can represent in your work.

Oral presentation

Your oral presentation will be evaluated on the same criteria as your report, but also on your ability to answer technical questions and on the quality of the answers provided.
We remind you that an oral presentation does not consist of reading a sheet of paper on which you have written a script. During the oral presentation, you can bring documents to help you, but it is out of the question that it turns into a reading session.

Specific instructions for students enrolled in train-linked training

Extra work will be required for students enrolled in train-linked training. The exact content will be discussed with the two teachers.

Practical instructions on your submission

The work you have to return must be composed of two items:

- a PDF document containing your report (≈ 25 to 35 pages)
- the scripts and data used.

You will place these two items in a zipped directory named as follows: the name of the first group member, followed by the name of the second. For example: Name_1-Name_2.zip.
Note: do NOT send your work by email.

Due date of the report

The files must be returned on the AMeTICE Moodle platform no later than January 21th 2023, at 11:59 p.m. Submissions will not be possible after this date.

Date of the oral presentation

The date of the oral will be communicated to you later. We will organise this by the end of January.